Early key takeaways of Peishan Li (Team 4)

Dataset overview

- Dataset: ‘luad_tcga_clinical_data.tsv’ (Non-Small Cell Lung Cancer/ Lung Adenocarcinoma)
- Reason of selecting this dataset: Interest in the topic; large sample size with minimal missing values
- Descriptive statistics:

```
df.describe()
```

	Diagnosis Age	Last Alive Less Initial Pathologic Diagnosis Date Calculated Day Value	Fraction Genome Altered	Year Cancer Initial Diagnosis	Overall Survival (Months)	Number of Samples Per Patient	Sample type id	Patient Smoking History Category
count	505.000000	505.0	518.000000	514.000000	515.000000	586.000000	586.000000	510.000000
mean	65.338614	0.0	0.266012	2008.383268	29.739903	1.006826	1.003413	2.813725
std	10.004275	0.0	0.190427	4.163682	29.275243	0.082407	0.058371	1.081460
min	33.000000	0.0	0.000000	1991.000000	0.000000	1.000000	1.000000	1.000000
25%	59.000000	0.0	0.104600	2007.000000	13.685000	1.000000	1.000000	2.000000
50%	66.000000	0.0	0.237800	2010.000000	21.580000	1.000000	1.000000	3.000000
75%	73.000000	0.0	0.400400	2011.000000	37.305000	1.000000	1.000000	4.000000
max	88.000000	0.0	0.801300	2013.000000	238.110000	2.000000	2.000000	5.000000

586 rows in total

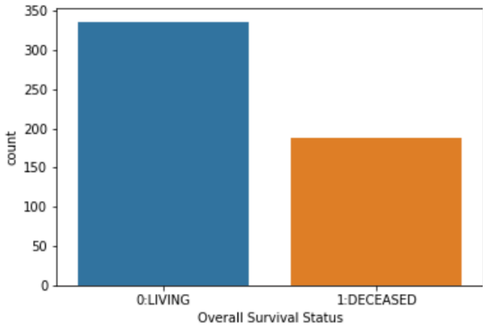
```
missing=df.isnull().sum()
missing=missing.to_frame('Missing count').reset_index()
missing=missing.rename(columns={'index': 'Variable name'})
missing
var_list=missing[missing['Missing count']<100]['Variable name'].tolist()
df=df[var_list]
```

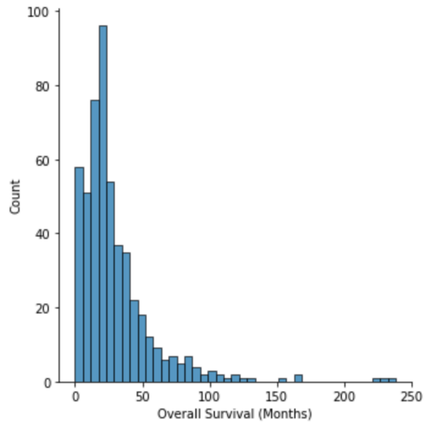
Leave out variables with more than 100 missing values

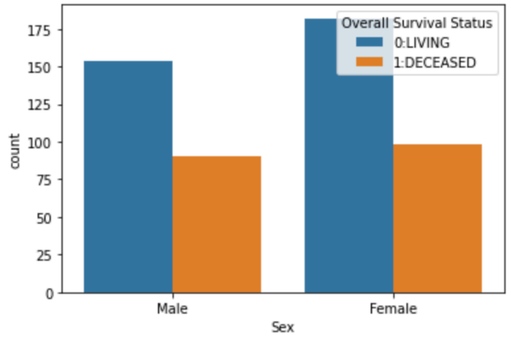
```
Study ID                                                                        object
Patient ID                                                                      object
Sample ID                                                                       object
Diagnosis Age                                                                  float64
American Joint Committee on Cancer Metastasis Stage Code                        object
Neoplasm Disease Lymph Node Stage American Joint Committee on Cancer Code       object
Neoplasm Disease Stage American Joint Committee on Cancer Code                  object
American Joint Committee on Cancer Publication Version Type                     object
American Joint Committee on Cancer Tumor Stage Code                             object
Cancer Type                                                                     object
Cancer Type Detailed                                                            object
Last Alive Less Initial Pathologic Diagnosis Date Calculated Day Value         float64
Form completion date                                                            object
Fraction Genome Altered                                                        float64
Neoplasm Histologic Type Name                                                   object
Neoadjuvant Therapy Type Administered Prior To Resection Text                   object
Prior Cancer Diagnosis Occurence                                                object
ICD-10 Classification                                                           object
International Classification of Diseases for Oncology, Third Edition ICD-O-3 Histology Code    object
International Classification of Diseases for Oncology, Third Edition ICD-O-3 Site Code        object
Informed consent verified                                                       object
Year Cancer Initial Diagnosis                                                  float64
Is FFPE                                                                         object
Oncotree Code                                                                   object
Overall Survival (Months)                                                      float64
Overall Survival Status                                                         object
Other Patient ID                                                                object
Other Sample ID                                                                 object
Pathology Report File Name                                                      object
Pathology report uuid                                                           object
Patient Primary Tumor Site                                                      object
Tissue Prospective Collection Indicator                                         object
Tissue Retrospective Collection Indicator                                       object
Number of Samples Per Patient                                                    int64
Sample Type                                                                     object
Sample type id                                                                   int64
Sex                                                                             object
Somatic Status                                                                  object
Tissue Source Site                                                              object
Patient Smoking History Category                                               float64
Tumor Site                                                                      object
Vial number                                                                     object
dtype: object
```

Exploratory data analysis and interpretations

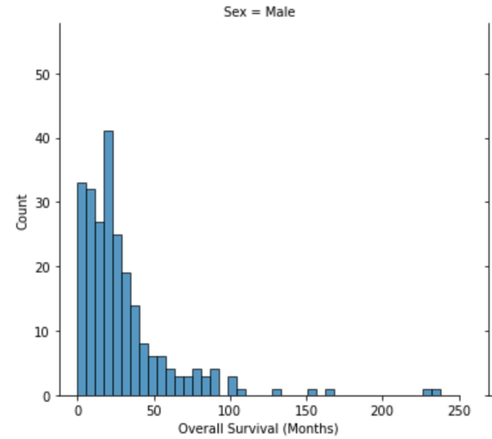
- **Overall survival status**: 336 Living VS 188 Deceased
- **Overall survival months**: Mostly distributed between 0-100 months;
- with a peak at around 30-40 months of survival
- **Lung Adenocarcinoma survival status with gender**:
- Female survival rate as 65% and male 63.11% in general
- Women have higher peak at around 30 months of survival than men;
- Women have more cases of survival between 100-150 months

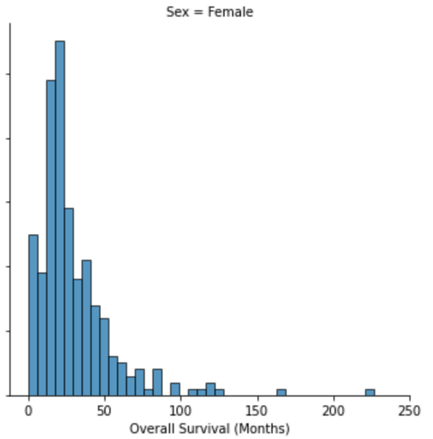






Sex	Overall Survival Status	Count
Female	0:LIVING	182
	1:DECEASED	98
Male	0:LIVING	154
	1:DECEASED	90



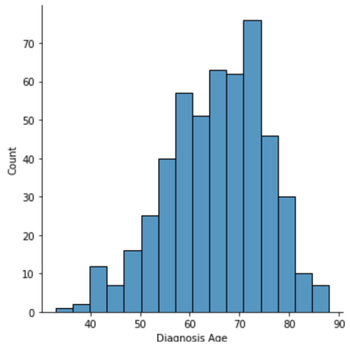


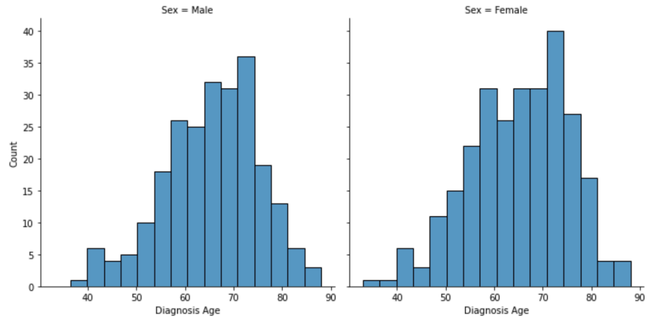
Exploratory data analysis and interpretations (Continued)

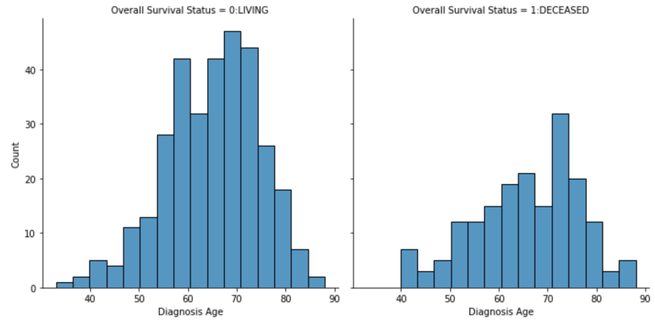
- **Lung Adenocarcinoma survival status with diagnosis age:**

- See distribution plots of diagnosis age, and its segmentation based on sex and survival status: For both sex, and for living and deceased, the mean of diagnosis age is around 65;

- Indicating it might not be an influencing factor of survival status

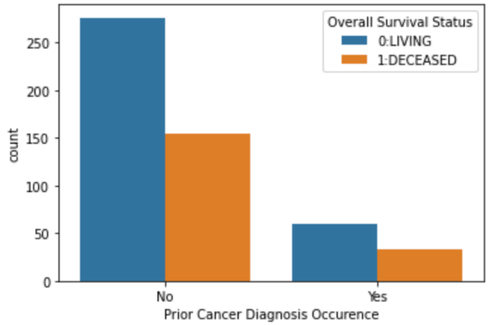






- **Lung Adenocarcinoma survival status with prior diagnosis:**

Prior Cancer Diagnosis Occurence	Overall Survival Status	Count
No	0:LIVING	276
	1:DECEASED	155
Yes	0:LIVING	60
	1:DECEASED	33



- The survival rate is both around 64% for people with/without prior diagnosis

- **Lung Adenocarcinoma survival status with primary tumor site:**

- The survival rate is highest if tumor is detected at R-middle, and lowest at R-lower

Patient Primary Tumor Site	Overall Survival Status	Count
L-Lower	0:LIVING	49
	1:DECEASED	30
L-Upper	0:LIVING	80
	1:DECEASED	45
Other (please specify)	0:LIVING	3
	1:DECEASED	1
R-Lower	0:LIVING	59
	1:DECEASED	40
R-Middle	0:LIVING	16
	1:DECEASED	5
R-Upper	0:LIVING	124
	1:DECEASED	61